D154–D158 Nucleic Acids Research, 2008, Vol. 36, Database issue Published online 8 November 2007
doi:10.1093/nar/gkm952

# miRBase: tools for microRNA genomics

Sam Griffiths-Jones[1],*, Harpreet Kaur Saini[2], Stijn van Dongen[2] and Anton J. Enright[2]

[1]Faculty of Life Sciences, University of Manchester, Michael Smith Building, Oxford Road, Manchester, M13 9PT and [2]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, Hinxton, UK

Received September 14, 2007; Revised October 10, 2007; Accepted October 16, 2007

## ABSTRACT

**miRBase is the central online repository for microRNA (miRNA) nomenclature, sequence data, annotation and target prediction. The current release (10.0) contains 5071 miRNA loci from 58 species, expressing 5922 distinct mature miRNA sequences: a growth of over 2000 sequences in the past 2 years. miRBase provides a range of data to facilitate studies of miRNA genomics: all miRNAs are mapped to their genomic coordinates. Clusters of miRNA sequences in the genome are highlighted, and can be defined and retrieved with any inter-miRNA distance. The overlap of miRNA sequences with annotated transcripts, both protein- and non-coding, are described. Finally, graphical views of the locations of a wide range of genomic features in model organisms allow for the first time the prediction of the likely boundaries of many miRNA primary transcripts. miRBase is available at http://microrna.sanger.ac.uk/.**

## INTRODUCTION

MicroRNAs (miRNAs) are short RNA sequences expressed from longer transcripts encoded in animal, plant and virus genomes, and recently discovered in a single-celled eukaryote (1,2). miRNAs regulate the expression of target genes by binding to complementary sites in their transcripts to cause translational repression or transcript degradation (3). Translational repression is thought to be the primary mechanism for imperfect target duplexes in animals, with transcript degradation the dominant mechanism for largely perfect matches found throughout plant target transcripts. miRNAs have been implicated in processes and pathways such as development, cell proliferation, apoptosis, metabolism and morphogenesis, and in diseases including cancer (4,5).

miRBase is the primary repository and database resource for miRNA data. The database has three main functions:

(i) miRBase::Registry provides a confidential service for the independent assignment of names to novel miRNA genes prior to their publication in peer-reviewed journals. Over 70 publications describing novel miRNA genes have made use of this service, and registration is a requirement of many journals.
(ii) miRBase::Sequences provides miRNA sequence data, annotation, references and links to other resources for all published miRNAs. The database (release 10.0) contains over 5000 sequences from 58 species.
(iii) miRBase::Targets provides an automated pipeline for the prediction of targets for all published animal miRNAs. The current release of the database (v5) predicts targets in over 500 000 transcripts for all miRNAs in 24 species. The target prediction pipeline and algorithms have been described elsewhere (6,7).

The miRNA nomenclature scheme has been presented and discussed previously (6,8,9). Novel miRNAs require cloning or expression evidence, and should be submitted only after a manuscript describing their identification is accepted for publication. Assigned names should then be incorporated into the final version of the manuscript prior to publication. Obvious homologues of miRNAs validated in closely related species need not be experimentally verified and may be submitted at any time. Primary features of the nomenclature scheme are:

(i) The miRNA name contains a three or four letter species prefix and a numeric suffix (e.g. hsa-mir-212).
(ii) A mature miRNA sequence may be predicted to be expressed from more than one hairpin precursor locus, denoted with further numeric suffixes (e.g. dme-mir-6-1 and dme-mir-6-2).
(iii) Related hairpin loci expressing related mature miRNA sequences have lettered suffixes (e.g. mmu-mir-181a and mmu-mir-181b).

*To whom correspondence should be addressed. Tel: +44 161 2755673; Fax: +44 161 2755082; Email: sam.griffiths-jones@manchester.ac.uk

(iv) Plant miRNA genes are given names of the form ath-MIR166a. Lettered suffixes describe distinct loci expressing all related mature miRNAs; numeric suffixes are not used.

(v) Viral miRNA names conventionally relate to the locus from which the miRNA derives (e.g. ebv-mir-BART1 from the Epstein Barr virus BART locus).

However, it is important to note that a short name cannot always encode complex information such as orthology and paralogy relationships. In some cases, the short name is a pragmatic choice that is the most consistent of conflicting representations of these sequence relationships. While the names provide a guide of family and function, they should not therefore be relied upon to confer any complex meaning. Instead, dedicated fields in the database provide information about gene and mature miRNA sequence families.

The published miRNA literature is huge. Readers are referred to a number of comprehensive reviews of miRNA structure, biogenesis and function (4,10–12). Here, we focus on specific issues and points of interest with respect to the provision of miRNA data in the miRBase database.

## miRBase DATA AND UPDATES

### How many miRNA genes?

The number of miRNA hairpin loci in the miRBase database continues to grow rapidly, from 2909 in 36 genomes (June 2005, release 7.0) to 5071 in 58 genomes (August 2007, release 10.0) in the past 2 years. The number of miRNAs in a genome has been the subject of much discussion in the literature. Early estimates of the number of miRNAs in the worm and human genomes were put at 123 and 255, respectively (13,14). However, these estimates were based largely on conservation studies. It is now clear that many miRNAs may be clade- or even organism-specific. A number of recent large-scale studies have lifted the number of miRNA loci known in human to 533 (Table 1) (15–17), around 60% of which are obviously conserved in mouse (miRBase release 10.0).

### miR and miR* sequences

The 5071 miRNA hairpin loci in the database express 4922 dominant mature miRNA (miR) products (Table 1). In many cases, deep sequencing technologies have detected large numbers of miR* sequences—biogenesis byproducts that are often detected at very low levels and are likely non-functional. Starting in miRBase release 10.0, mature miR and miR* sequences are better distinguished in the database, and distributed in separate release files. In many cases, mature miRNAs from both 5′ and 3′ arms of the hairpin precursor are frequently identified, suggesting that both may be functional, or there is insufficient data to determine the predominant product. Such miRNAs are given names of the form hsa-miR-140-5p and hsa-miR-140-3p, and both are retained in the miR set. Often, subsequent improved data allow one product to be chosen and annotated as the dominant miR. Recent data updates have occasionally caused the annotation of a miR and miR* pair to be reversed.

### Variable ends

Increasingly deep and comprehensive cloning and sequencing studies identify many mature miRNAs with variable 3’ (and, to a lesser extent, 5’) ends [see for example (17)]. The miRNAs in the database currently represent the consensus of the most dominantly expressed sequence. As more data become available, the ends of mature miRNAs in the database will be adjusted to reflect the most up-to-date consensus information. We also aim to provide specific data on the distribution of ends in future releases. All changes in name and sequence between releases are specifically described in the diff file on the FTP site, along with all data from previous releases.

### Experimental support

Usually the only available experimental data supports the mature miRNAs—hairpin precursors are very rarely experimentally validated. Rather, the precursors are the result of computational prediction of hairpin structures that include the mature miRNA. When a number of loci include the same mature miRNA, we cannot usually say with confidence which loci are actually expressed. In addition, the extents of the hairpins depicted in the database are somewhat arbitrary—the approximate extent of the predicted hairpin structure is shown. Formally, this includes the true precursor (the product of DROSHA cleavage) and a

**Table 1.** The number of published hairpin precursor and mature miRNA sequences in selected model organisms

| | Hairpin precursor loci | | | Mature miR sequences[a] | |
|---|---|---|---|---|---|
| | Total number | Clustered ≤10 kb from another miRNA | Overlap annotated transcripts | Distinct forms | Experimentally verified |
| *Homo sapiens* | 533 | 190 (36%) | 267 (50%) | 555 | 546 (98%) |
| *Mus musculus* | 442 | 199 (45%) | 174 (39%) | 461 | 455 (99%) |
| *Danio rerio* | 337 | 151 (34%) | 41 (12%) | 193 | 183 (95%) |
| *Caenorhabditis elegans* | 135 | 34 (25%) | 23 (17%) | 135 | 135 (100%) |
| *Drosophila melanogaster* | 93 | 34 (36%) | 36 (39%) | 88 | 85 (97%) |
| *Arabidopsis thaliana* | 184 | 19 (10%) | 16 (9%) | 199 | 199 (100%) |
| *Populus trichocarpa* | 215 | 42 (20%) | 9 (4%) | 215 | 55 (26%) |

[a]miR* sequences are excluded from the mature miRNA count.

small amount of flanking sequence. Future developments will include the provision to retrieve the precursor with user-defined lengths of flanking sequence. About 3685 of 5922 mature miRNA products in the database are validated experimentally in the originating organism—the remainders are obvious homologues of validated miRNAs from a related species (Table 1). The 'evidence' field describes the origin of each sequence in the database.

### miRBase::Targets

The miRBase::Targets database uses the miRanda algorithm (7) to predict targets in untranslated regions (UTRs) of 37 animal genomes from Ensembl (18). The quality of the predictions has recently benefited from significantly improved 3′UTR information, based on DITAG and 5′CAGE data, available from Ensembl. The number of human and mouse transcripts without an experimentally supported 3′UTR (for which we search a region 2 kb downstream) has therefore dropped significantly in the latest release (v5). A number of validated miR/target pairs are shown to have mismatches in the so-called 'seed' region (19). The miRBase/miRanda pipeline is therefore not constrained by the requirement for exact 'seed' matches. Recent papers have also highlighted the importance of secondary features for miRNA/target recognition, such as sequence accessibility, AU bias and UTR position (20,21). We intend to incorporate these features into the miRBase::Target prediction pipeline over the coming 12 months. In addition, links are provided to other target prediction sites and algorithms, and to the TarBase database of experimentally supported targets (22).

## miRBase GENOMICS

Recently, we have focused on the provision of tools to distribute miRNA genomic information.

### Genomic coordinates

Where an assembled genome sequence is available, coordinates of all miRNAs are provided: in summary tables for each organism and miRNA family, on each miRNA entry page, and for bulk download in GFF format. Links are provided from each coordinate to the appropriate genome browsers.

### miRNA gene context

40–70% of vertebrate miRNAs appear to be expressed from introns of protein- and non-coding transcripts (Table 1) (23). In worms and flies, intronic miRNAs are less common (15% and 39%, respectively, in protein-coding genes), and only 5–10% of *Arabidopsis* miRNAs overlap annotated transcripts. For all animals with Ensembl-annotated genome assemblies, we provide a list of transcripts overlapping each miRNA, with overlap type (intron, exon and UTR), and sense (forward and reverse strands).

### Clustered miRNAs

miRNAs are often clustered close together in the genome. This clustering has been suggested as evidence that >1 miRNA may be expressed from the same primary miRNA transcript (pri-miRNA). Furthermore, known 'polycistronic' miRNA transcripts are shown to be long: up to tens of kilobases in mammals. Over 40% of human miRNAs, over 30% of worm and fly miRNAs and only around 10% of *Arabidopsis* miRNAs are within 10 kb of another miRNA (Table 1). miRBase provides a list of clustered miRNAs on each applicable entry page. In addition, a new search facility allows the user to retrieve clusters of miRNAs in any organism separated by any choice of distance.

### Genomic features

While the mapping of mature and hairpin miRNA sequences to assembled genomes is readily available in miRBase, the extents of only very few primary miRNA transcripts (pri-miRNA) are determined and annotated. For intronic miRNAs, the pri-miRNA is assumed to be the protein- (or non-)coding host transcript. Information about the extents of intergenic pri-miRNAs can be inferred from collective analysis of genomic features such as transcription start sites (TSS), CpG islands, EST and cDNA overlap, DITAG and 5′CAGE data, transcription factor binding sites (TFBS) and polyadenylation site predictions (polyA). A detailed analysis of these data suggest that pri-miRNA transcripts vary in length from a few hundreds of bases up to tens of kilobases (24). We have recently developed a tool to visualize the relative positions of these predictions and mappings with respect to annotated miRNA genes and clusters. Careful inspection of these data allows the prediction of the 5′ and 3′ boundaries of a significant number of putative pri-miRNAs. For example, Figure 1 shows TSSs, CpG island, ESTs, cDNAs, DITAG (172B22 and 172B221) and polyA site predictions surrounding mmu-mir-135b on mouse chromosome 1, which support a primary transcript of length around 15 kb with 5′ and 3′ ends ~7–8 kb upstream and downstream of the miRNA. Links from each miRNA entry page provide a tabulated list of features overlapping flanking regions of the miRNA with their corresponding coordinates and scores, and a graphical view of the features present in the miRNA gene neighbourhood (as in Figure 1). These views are currently available for human, mouse, rat, worm and fly miRNAs, and will be extended to other organisms in the future. For human, mouse and rat genomes, TSSs are predicted using the Eponine-TSS software (25) at a threshold of 0.990. *Drosophila* TSS predictions, together with CpG islands, ESTs, cDNAs, repeats and DITAGs for all species are obtained from Ensembl. TFBSs in the flanking regions of human miRNAs are obtained from the conserved TFBS track of the UCSC genome browser (26). Other TFBS data are imported from the regulatory features track of Ensembl. PolyA signals are predicted in-house using the DNAFSMiner method (27) with a cutoff score of 0.6. The 'Genomics' section of the

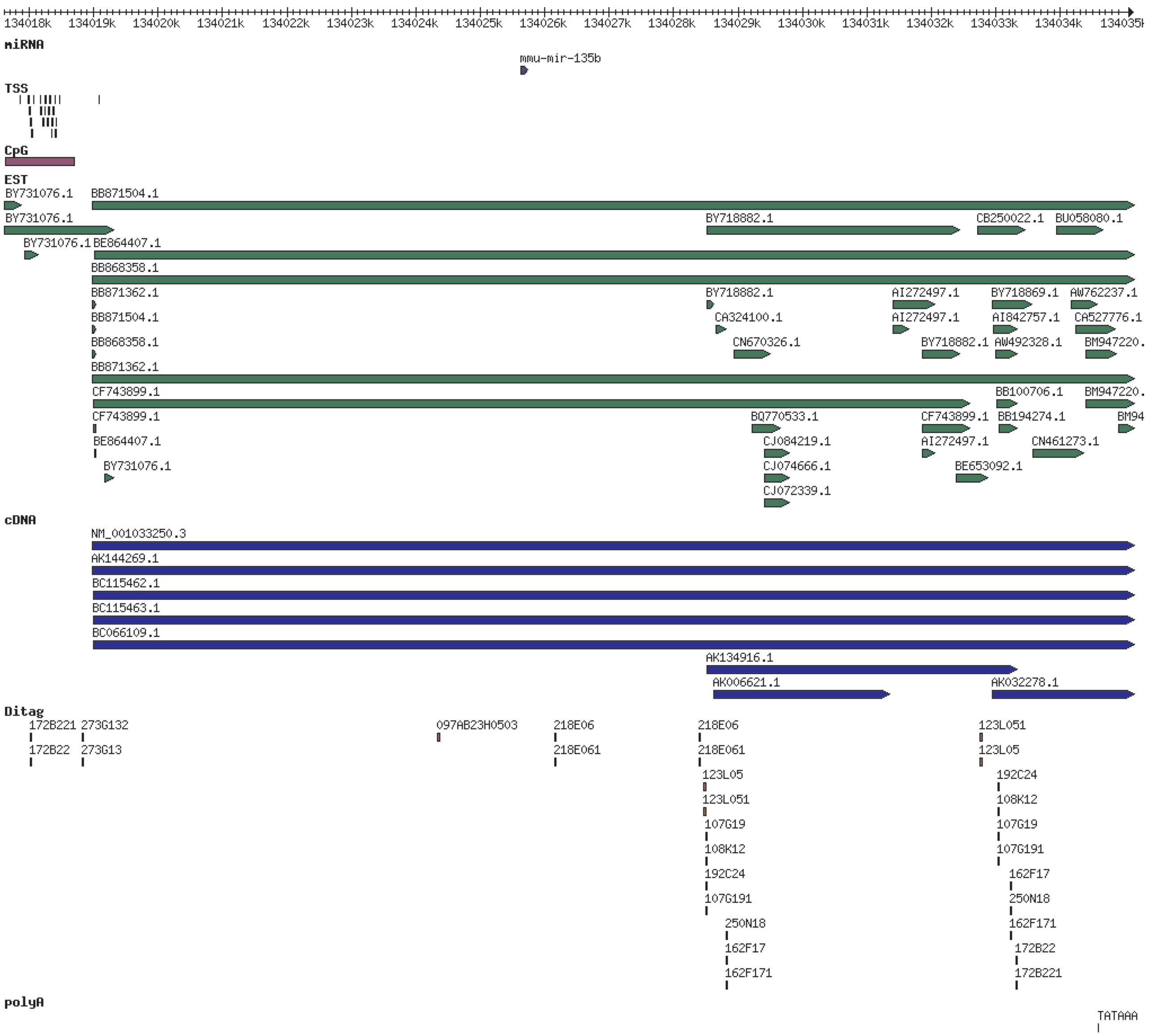

**Figure 1.** miRBase view of the distribution of genomic features around mmu-mir-135b on mouse chromosome 1, showing TSS, CpG island, EST, cDNA, DITAG (172B221 and 172B22) and polyA site support for a 15 kb primary transcript.

miRBase site allows the user to specify flanking and clustering distances, and the range of features desired.

## AVAILABILITY

miRBase is available on the web at http://microrna.sanger.ac.uk/. All data are available for download from the FTP site (ftp://ftp.sanger.ac.uk/pub/mirbase/) in a variety of formats including FASTA sequences and MYSQL relational database dumps.

## ACKNOWLEDGEMENTS

S.G.-J. is funded by the University of Manchester. H.K.S. holds a GlaxoSmithKline postdoctoral fellowship, and work at the Sanger Institute is funded by the Wellcome Trust. Funding to pay the Open Access publication charges for this article was provided by the University of Manchester.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Molnár,A., Schwach,F., Studholme,D.J., Thuenemann,E.C. and Baulcombe,D.C. (2007) miRNAs control gene expression in the single-cell alga *Chlamydomonas reinhardtii*. *Nature*, **447**, 1126–1129.
2. Zhao,T., Li,G., Mi,S., Li,S., Hannon,G.J., Wang,X.J. and Qi,Y. (2007) A complex system of small RNAs in the unicellular green alga *Chlamydomonas reinhardtii*. *Genes Dev.*, **21**, 1190–1203.
3. Pillai,R.S., Bhattacharyya,S.N. and Filipowicz,W. (2007) Repression of protein synthesis by miRNAs: how many mechanisms? *Trends Cell Biol.*, **17**, 118–126.
4. Kloosterman,W.P. and Plasterk,R.H. (2006) The diverse functions of microRNAs in animal development and disease. *Dev. Cell*, **11**, 441–450.

5. Garzon,R., Fabbri,M., Cimmino,A., Calin,G.A. and Croce,C.M. (2006) MicroRNA expression and function in cancer. *Trends Mol. Med.*, **12**, 580–587.
6. Griffiths-Jones,S., Grocock,R.J., van Dongen,S., Bateman,A. and Enright,A.J. (2006) miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.*, **34**, D140–D144.
7. John,B., Enright,A.J., Aravin,A., Tuschl,T., Sander,C. and Marks,D.S. (2004) Human microRNA targets. *PLoS Biol.*, **2**, e363.
8. Ambros,V., Bartel,B., Bartel,D.P., Burge,C.B., Carrington,J.C., Chen,X., Dreyfuss,G., Eddy,S.R., Griffiths-Jones,S. *et al.* (2003) A uniform system for microRNA annotation. *RNA*, **9**, 277–279.
9. Griffiths-Jones,S. (2004) The microRNA registry. *Nucleic Acids Res.*, **32**, D109–D111.
10. Kim,V.N. and Nam,J.W. (2006) Genomics of microRNA. *Trends Genet.*, **22**, 165–173.
11. Kim,V.N. (2005) MicroRNA biogenesis: coordinated cropping and dicing. *Nat. Rev. Mol. Cell. Biol.*, **6**, 376–385.
12. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
13. Lim,L.P., Lau,N.C., Weinstein,E.G., Abdelhakim,A., Yekta,S., Rhoades,M.W., Burge,C.B. and Bartel,D.P. (2003) The microRNAs of *Caenorhabditis elegans*. *Genes Dev.*, **17**, 991–1008.
14. Lim,L.P., Glasner,M.E., Yekta,S., Burge,C.B. and Bartel,D.P. (2003) Vertebrate microRNA genes. *Science*, **299**, 1540.
15. Cummins,J.M., He,Y., Leary,R.J., Pagliarini,R., Diaz,L.A., Sjoblom,T., Barad,O., Bentwich,Z., Szafranska,A.E. *et al.* (2006) The colorectal microRNAome. *Proc. Natl Acad. Sci. USA*, **103**, 3687–3692.
16. Berezikov,E., van Tetering,G., Verheul,M., van de Belt,J., van Laake,L., Vos,J., Verloop,R., van de Wetering,M., Guryev,V. *et al.* (2006) Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. *Genome Res.*, **16**, 1289–1298.
17. Landgraf,P., Rusu,M., Sheridan,R., Sewer,A., Iovino,N., Aravin,A., Pfeffer,S., Rice,A., Kamphorst,A.O. *et al.* (2007) A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell*, **129**, 1401–1414.
18. Hubbard,T.J., Aken,B.L., Beal,K., Ballester,B., Caccamo,M., Chen,Y., Clarke,L., Coates,G., Cunningham,F. *et al.* (2007) Ensembl 2007. *Nucleic Acids Res.*, **35**, D610–D617.
19. Didiano,D. and Hobert,O. (2006) Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. *Nat. Struct. Mol. Biol.*, **13**, 849–845.
20. Long,D., Lee,R., Williams,P., Chan,C.Y., Ambros,V. and Ding,Y. (2007) Potent effect of target structure on microRNA function. *Nat. Struct. Mol. Biol.*, **14**, 287–294.
21. Grimson,A., Farh,K.K., Johnston,W.K., Garrett-Engele,P., Lim,L.P. and Bartel,DP. (2007) MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol. Cell*, **27**, 91–105.
22. Sethupathy,P., Corda,B. and Hatzigeorgiou,A.G. (2006) TarBase: a comprehensive database of experimentally supported animal microRNA targets. *RNA*, **12**, 192–197.
23. Rodriguez,A., Griffiths-Jones,S., Ashurst,J.L. and Bradley,A. (2004) Identification of mammalian microRNA host genes and transcription units. *Genome Res.*, **14**, 1902–1910.
24. Saini,H.K., Griffiths-Jones,S. and Enright,A.J. (2007) Genomic analysis of human microRNA transcripts. *Proc. Natl Acad. Sci. USA*, **104**, 17719–17724.
25. Down,T.A. and Hubbard,T.J. (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. *Genome Res.*, **12**, 458–461.
26. Kuhn,R.M., Karolchik,D., Zweig,A.S., Trumbower,H., Thomas,D.J., Thakkapallayil,A., Sugnet,C.W., Stanke,M., Smith,K.E. *et al.* (2007) The UCSC genome browser database: update 2007. *Nucleic Acids Res.*, **35**, D668–D673.
27. Liu,H., Han,H., Li,J. and Wong,L. (2005) DNAFSMiner: a web-based software toolbox to recognize two types of functional sites in DNA sequences. *Bioinformatics*, **21**, 671–673.